

THE IMPORTANCE OF THE RANGE PARAMETER FOR ESTIMATION AND PREDICTION IN GEOSTATISTICS

BY CARI G. KAUFMAN AND BENJAMIN A. SHABY*

University of California, Berkeley and Duke University

Two canonical problems in geostatistics are estimating the parameters in a specified family of stochastic process models and predicting the process at new locations. A number of asymptotic results for these problems over a fixed spatial domain indicate that, for a Gaussian process with Matérn covariance function, one can fix the range parameter controlling the rate of decay of the process and obtain results that are asymptotically equivalent to the case that the range parameter is known. We discuss why these results do not always provide the appropriate intuition for finite samples. Moreover, we prove that a number of these asymptotic results may be extended to the case that the variance and range parameters are jointly estimated via maximum likelihood or maximum tapered likelihood. Our simulation results show that performance on a variety of metrics is improved and asymptotic approximations are applicable for smaller sample sizes when the range parameter is estimated. These effects are particularly apparent when the process is mean square differentiable or the effective range of spatial correlation is small.

1. Introduction. The analysis of point-referenced spatial data, often referred to as geostatistics, relies almost exclusively on a single construct: the second-order stationary Gaussian process with a parametric mean and covariance function. Exceptions may be found, some of them notable, but in almost all elaborate hierarchical or nonstationary models in the literature, one can find structures built from stationary Gaussian processes.

Given the prominent role of the stationary Gaussian process, it is perhaps surprising that the theoretical properties of inference under this model remain incompletely understood. Consider a canonical problem in geostatistics, that of predicting the value of a spatial process at locations not contained in the dataset. The steps of the analysis using a Gaussian process can be characterized as follows: choose a parametric mean and covariance function, estimate the model parameters from the data, then plug the estimated parameters back into the model to make predictions and compute

*Supported by the National Science Foundation under Grant DMS-0914906.

AMS 2000 subject classifications: Primary 62M30; secondary 62F12, 62M20

Keywords and phrases: Gaussian processes, covariance estimation, infill asymptotics, Matérn covariance

standard errors at unobserved locations. Discussion of these related steps can be found in most spatial statistics textbooks, but see for example [Cressie \(1993, Chapters 2 and 3\)](#), or see [Banerjee, Carlin and Gelfand \(2004\)](#) for a Bayesian treatment. [Stein \(2010\)](#) gives a succinct overview of asymptotic issues for both estimation and prediction.

Covariance model choice, estimation, and prediction have each received considerable theoretical attention, although often in isolation from one another. A thorough treatment of the problem of model choice may be found in [Stein \(1999\)](#). In this work and elsewhere, Stein makes a compelling case for using the Matérn covariance model for the Gaussian process $\{Z(s), s \in D \subseteq \mathbb{R}^d\}$, with

(1)

$$\text{Cov}(Z(s_i), Z(s_j)) = \sigma^2 K(s_i - s_j; \rho, \nu) = \frac{\sigma^2 (\|s_i - s_j\|/\rho)^\nu}{\Gamma(\nu) 2^{\nu-1}} \mathcal{K}_\nu(\|s_i - s_j\|/\rho),$$

for $\sigma^2, \rho, \nu > 0$, where \mathcal{K}_ν is the modified Bessel function of the second kind of order ν ([Abramowitz and Stegun, 1967, Section 9.6](#)). The range parameter ρ controls the rate of decay with distance, with larger values of ρ corresponding to more highly correlated observations. This model is particularly attractive because of its flexibility in representing the smoothness of the Gaussian process, with any number of mean square derivatives being possible, according to the value of ν ([Stein, 1999](#)).

[Zhang \(2004\)](#) provides influential results concerning the consistency (and *inconsistency*) of parameter estimates for the Matérn model under infill, or fixed-domain asymptotics. Infill asymptotics requires that the sampling domain be fixed as the number, and hence density, of observations increases to infinity. These results follow from a more fundamental result in [Zhang \(2004\)](#) concerning equivalence (mutual absolute continuity) of Gaussian measures on bounded domains.

A theoretical treatment of the third component of geostatistical analyses, prediction and corresponding standard error estimation, has been developed in a series of works by [Stein \(1988, 1990, 1993, 1999\)](#). These works provide conditions under which predictions using a mis-specified covariance function are asymptotically efficient and associated standard error estimates converge almost surely to their true values under infill asymptotics. One such condition is that the mis-specified covariance be chosen so that the resulting Gaussian measure and the true one are equivalent, providing a link to the results in [Zhang \(2004\)](#). However, as we will discuss in [Section 2.2](#), the nature of that link has sometimes been misinterpreted.

Like most of the aforementioned works, we focus on the isotropic d -dimensional Matérn covariance model. Our goal is to provide a holistic ac-

count of the asymptotic properties of the canonical geostatistical analysis under the Matérn model, from parameter estimation and inference to prediction. This account includes the re-statement of key results, as well as the introduction of new results to fill important gaps.

We devote particular attention to the range parameter ρ . One may detect in the recent literature a vein of reasoning that ρ is unimportant in practice. For example, [Zhang and Wang \(2010\)](#) make this argument in the context of prediction, and [Gneiting, Kleiber and Schlather \(2010\)](#) make a similar argument for using a single range parameter in a Matérn model for multivariate random fields. These authors borrow intuition from the asymptotic results of [Stein \(1988\)](#), [Zhang \(2004\)](#), and others, results that fix ρ at an arbitrary value, often for mathematical tractability. Each of these results presents some particular variation of the conclusion that fixing ρ at an incorrect value is asymptotically just as good as using the true value. However, as we will show via a simulation study in Section 3, this intuition cannot necessarily be transferred so readily to the finite sample case. Several of our new results indicate that the same asymptotic behaviors also hold for the case of estimated ρ , and we demonstrate in the simulation that the approximations provided by these results can be applicable for much smaller sample sizes than when ρ is fixed.

In the Section 2 we state and extend a number of important results for the Matérn model, moving from the case that ρ is fixed at an arbitrary value to the case that σ^2 and ρ are jointly estimated via maximum likelihood. We also review and extend asymptotic results for prediction, although this theory does not yet allow for estimated ρ . In Section 3 we carry out a simulation study that allows us to determine the cases in which estimation of the range parameter is important to performance and the cases in which it is not important. We show that instances in the literature where the range parameter was claimed not to matter correspond to particular scenarios and that this result does not hold more generally. In Section 4 we prove that similar arguments as in Section 2 for the maximum likelihood estimator can also be applied to an estimator maximizing an approximation using covariance tapering. We conclude in Section 5 with a discussion of the implications of these results and a number of ways in which assumptions we have made here might be relaxed.

2. Asymptotic Theory for Estimation and Prediction. We begin with some notation and assumptions that will be used in all our results unless specifically stated otherwise. Let $Z = \{Z(s), s \in D \subset \mathbb{R}^d\}$ be a stochastic process on a bounded domain D , with $d = 1, 2$, or 3 . Let $G(0, \sigma^2 K_\theta)$ denote

the mean zero stationary Gaussian measure for Z with marginal variance $\sigma^2 > 0$ and correlation function K_θ , depending on parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. For a particular sampling design $S_n = \{s_1, \dots, s_n \in D\}$ of distinct locations, we observe $Z_n = (Z(s_1), \dots, Z(s_n))^T$. Our tasks are to use Z_n to estimate σ^2 and θ and to predict $Z(s_0)$ for some location $s_0 \in D$, not in S_n . Our results concern the behavior of these estimators and predictors as we take an increasing number of observations in a bounded domain D , called “infill” or “fixed-domain” asymptotics.

Throughout, we focus our discussion on the Matérn covariance function (1), and we use $G(0, \sigma^2 K_{\rho, \nu})$ to denote a mean zero Gaussian measure with this covariance function. We also assume that the smoothness parameter ν is known. Although having the flexibility to estimate ν is desirable, results in this more general case have thus far been difficult to obtain under the fixed-domain asymptotics. Indeed, of the handful of results that have been obtained for estimators under this framework, we know of none that do make not the assumption that ν is known. Our focus is on the role played by the range parameter ρ in this model, namely to show that several important results that have been provable only in the case of fixing ρ at an arbitrary value can be extended to the case that ρ is estimated, and that it is often advantageous in practice to do so.

The reason that it is justifiable to fix ρ , at least in an asymptotic sense, follows from a property of the Matérn model shown by [Zhang \(2004\)](#). This result indicates that when the dimension $d \leq 3$, two Gaussian measures with the same ν but different values of ρ can in fact be equivalent, so that it is impossible to distinguish between them even when observing $Z(s)$ for all $s \in D$.

THEOREM 1 (Theorem 2 of [Zhang \(2004\)](#)). *For fixed $\nu > 0$, $G(0, \sigma_0^2 K_{\rho_0, \nu})$ and $G(0, \sigma_1^2 K_{\rho_1, \nu})$ are equivalent on bounded domains if and only if $\sigma_0^2 / \rho_0^{2\nu} = \sigma_1^2 / \rho_1^{2\nu}$.*

[Anderes \(2010\)](#) showed that this result does not hold for $d \geq 5$, and the case $d = 4$ is still open. However, the assumption that $d \leq 3$ seems adequate for most spatial applications.

The parameter $c = \sigma^2 / \rho^{2\nu}$ is what [Stein \(1999\)](#) calls a microergodic parameter. [Stein \(1999, page 175\)](#) gives a “reasonable conjecture” for microergodic parameters that we will show to be true in this case. He suggests reparameterizing into microergodic and non-microergodic components of the parameter vector, which we could choose to define as c and ρ , respectively. The conjecture is that if all model parameters are estimated by maximum likelihood, the asymptotic behavior of the MLE for the microergodic param-

eter, here c , is the same as if the non-ergodic component, here ρ , were known. In the next section, we outline existing results that concern the asymptotic behavior for the MLE for c when ρ is fixed at an arbitrary value, and we extend them to the case that ρ is estimated. This distinction is important in practice in a number of cases, as we highlight in the simulation study of Section 3. Section 2.2 turns to asymptotic theory for the prediction problem. The role of the microergodic parameter is critical here as well.

2.1. Estimation of Covariance Parameters. Theorem 1 has an immediate and important corollary for estimation.

COROLLARY 1 (Corollary 1 of Zhang (2004)). *Let S_n be an increasing sequence of subsets of D . There do not exist consistent estimators of σ^2 or ρ based on the corresponding sequence of observation vectors Z_n .*

It is important to note what Corollary 1 does and does not imply. It states that as $n \rightarrow \infty$, the sampling distributions of estimators of σ^2 and ρ , including, for example, the maximum likelihood estimators, do not concentrate their mass about the true parameter values. However, this does not mean that the data contain no information about σ^2 and ρ individually. Indeed, in simulation studies we observe that sampling distributions for the maximum likelihood estimators can in many cases be quite concentrated about the true values, even as we know these distributions will not become ever more concentrated as n increases (Zhang, 2004; Kaufman, 2006). Some intuition behind this can be given by appealing to another asymptotic framework, that of increasing the domain of observations. Mardia and Marshall (1984) give regularity conditions under which the maximum likelihood estimators for all model parameters are consistent and asymptotically normally distributed under the increasing domain framework. Any finite set of observation locations could conceivably be a member in a sequence under either the fixed-domain or increasing-domain asymptotic framework. Zhang and Zimmerman (2005) suggest choosing between the asymptotic frameworks based on the degree to which asymptotic approximations hold for finite samples. They also note that the increasing domain framework can be mimicked by fixing the domain but decreasing the range parameter. Therefore, it is not surprising that when the true range parameter is small relative to the sampling domain, it can be well estimated from data.

The likelihood function for σ^2 and ρ under the Matérn model with fixed $\nu > 0$ based on observations Z_n is

$$(2) \quad \mathcal{L}_n(\sigma^2, \rho) = (2\pi\sigma^2)^{-n/2} |\Gamma_n(\rho)|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} Z_n^T \Gamma_n(\rho)^{-1} Z_n \right\},$$

where $\Gamma_n(\rho)$ is the matrix with entries $K(s_i - s_j; \rho, \nu)$ ($i, j = 1, \dots, n$) for K defined as in (1). We consider two types of estimators obtained by maximizing (2). The first fixes $\hat{\rho}_n = \rho_1$ for all n and maximizes $\mathcal{L}_n(\sigma^2, \rho_1)$. The second maximizes (2) over both σ^2 and ρ . In either case, the estimator of σ^2 may be written as a function of the corresponding estimator of ρ . That is, we may write $\hat{\sigma}_n^2(\hat{\rho}_n) = \arg \max_{\sigma^2} \mathcal{L}_n(\sigma^2, \hat{\rho}_n) = Z_n^T \Gamma_n(\hat{\rho}_n)^{-1} Z_n / n$, where $\hat{\rho}_n$ is either ρ_1 or the value that maximizes the profile likelihood for ρ . In most cases the latter estimator is not available in closed form and must be found numerically.

We may likewise express the corresponding estimators of $c = \sigma^2 / \rho^{2\nu}$ as a function of $\hat{\rho}_n$, namely

$$(3) \quad \hat{c}_n(\hat{\rho}_n) = \hat{\sigma}_n^2(\hat{\rho}_n) / \hat{\rho}_n^{2\nu} = Z_n^T \Gamma_n(\hat{\rho}_n)^{-1} Z_n / (n \hat{\rho}_n^{2\nu})$$

The following result, taken from Theorem 3 of Zhang (2004) and Theorem 3 of Wang and Loh (2011), defines the asymptotic behavior of $\hat{c}_n(\rho_1)$ for an arbitrary fixed value $\rho_1 > 0$.

THEOREM 2. *Let S_n be an increasing sequence of subsets of D . Then as $n \rightarrow \infty$,*

1. $\hat{c}_n(\rho_1) \rightarrow \sigma_0^2 / \rho_0^{2\nu}$ almost surely, and
2. $\sqrt{n}(\hat{c}_n(\rho_1) - \sigma_0^2 / \rho_0^{2\nu}) \rightarrow N(0, 2(\sigma_0^2 / \rho_0^{2\nu})^2)$ in distribution

under $G(0, \sigma_0^2 K_{\rho_0, \nu})$.

Special cases of Theorem 2 were proven previously by Ying (1991), who treated the case that $\nu = 1/2$ with D an interval in \mathbb{R} , the Ornstein-Uhlenbeck process, and by Du, Zhang and Mandrekar (2009), who treated the more general case with $d = 1$.

A key contribution of the current paper is to show that Theorem 2 can be used as a stepping stone to proving that the maximum likelihood estimator $\hat{c}_n(\hat{\rho}_n)$ has exactly the same asymptotic behavior as does $\hat{c}_n(\rho_0)$. We make use of the following lemma, which shows that $\hat{c}_n(\hat{\rho}_n)$ is monotone when viewed as a function of $\hat{\rho}_n$.

LEMMA 1. *Let $S_n = \{s_1, s_2, \dots, s_n \in D \subseteq \mathbb{R}^d\}$ denote any set of observation locations in any dimension. Fix $\nu > 0$ and define $\Gamma_n(\rho)$ to be the matrix with entries $K(s_i - s_j; \rho, \nu)$ as in (1). Define $\hat{c}_n(\rho) = Z_n^T \Gamma_n(\rho)^{-1} Z_n / (n \rho^{2\nu})$. Then for any $0 < \rho_1 < \rho_2$, $\hat{c}_n(\rho_2) \leq \hat{c}_n(\rho_1)$ for any vector Z_n .*

PROOF. Let $0 < \rho_1 < \rho_2$. The difference

$$\hat{c}_n(\rho_1) - \hat{c}_n(\rho_2) = Z_n^T [\rho_1^{-2\nu} \Gamma_n(\rho_1)^{-1} - \rho_2^{-2\nu} \Gamma_n(\rho_2)^{-1}] Z_n / n$$

is non-negative for any Z_n if the matrix $A = \rho_1^{-2\nu} \Gamma_n(\rho_1)^{-1} - \rho_2^{-2\nu} \Gamma_n(\rho_2)^{-1}$ is positive semi-definite. By Corollary 7.7.4(a) of [Horn and Johnson \(1985, page 473\)](#), A is positive semi-definite if and only if the matrix $B = \rho_2^{2\nu} \Gamma_n(\rho_2) - \rho_1^{2\nu} \Gamma_n(\rho_1)$ is positive semi-definite. The entries of B may be expressed in terms of a function $K_B : \mathbb{R}^d \rightarrow \mathbb{R}$, with

$$B_{ij} = K_B(s_i - s_j) = \rho_2^{2\nu} K(\|s_i - s_j\|; \rho_2, \nu) - \rho_1^{2\nu} K(\|s_i - s_j\|; \rho_1, \nu),$$

and B is positive semi-definite if K_B is a positive definite function. Define

$$\begin{aligned} f_B(\omega) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega^T x} K_B(x) dx \\ (4) \quad &= \frac{1}{(2\pi)^d} \left[\rho_2^{2\nu} \int_{\mathbb{R}^d} e^{-i\omega^T x} K(x; \rho_2, \nu) dx - \rho_1^{2\nu} \int_{\mathbb{R}^d} e^{-i\omega^T x} K(x; \rho_1, \nu) dx \right]. \end{aligned}$$

Both integral terms in (4) are finite, with

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega^T x} K(x; \rho, \nu) dx = \frac{\Gamma(\nu + d/2)}{\pi^{d/2} \Gamma(\nu)} \rho^{-2\nu} (\rho^{-2} + \|\omega\|^2)^{-(\nu + d/2)},$$

the spectral density of the Matérn correlation function. Therefore,

$$f_B(\omega) = \frac{\Gamma(\nu + d/2)}{2^d \pi^{3d/2} \Gamma(\nu)} \left[(\rho_2^{-2} + \|\omega\|^2)^{-(\nu + d/2)} - (\rho_1^{-2} + \|\omega\|^2)^{-(\nu + d/2)} \right].$$

To show K_B is positive definite it suffices to show $f_B(\omega)$ is positive for all ω . This is clear because $0 < \rho_1 < \rho_2$. Therefore $\hat{c}_n(\rho_2) \leq \hat{c}_n(\rho_1)$ for any vector Z_n . \square

We can now make use of Theorem 2 in proving a more general result for the maximum likelihood estimator.

THEOREM 3. *Let S_n be an increasing sequence of subsets of D . Suppose $(\sigma_0^2, \rho_0)^T \in (0, \infty) \times [\rho_L, \rho_U]$, for any $0 < \rho_L < \rho_U < \infty$. Let $(\hat{\sigma}_n^2, \hat{\rho}_n)^T$ maximize (2) over $(0, \infty) \times [\rho_L, \rho_U]$. Then*

1. $\hat{\sigma}_n^2 / \hat{\rho}_n^{2\nu} \rightarrow \sigma_0^2 / \rho_0^{2\nu}$ almost surely, and
2. $\sqrt{n}(\hat{\sigma}_n^2 / \hat{\rho}_n^{2\nu} - \sigma_0^2 / \rho_0^{2\nu}) \rightarrow N(0, 2(\sigma_0^2 / \rho_0^{2\nu})^2)$ in distribution

under $G(0, \sigma_0^2 K_{\rho_0, \nu})$.

PROOF. By assumption, $\rho_L < \hat{\rho}_n < \rho_U$ for every n . Define two sequences, $\hat{c}_n(\rho_L)$ and $\hat{c}_n(\rho_U)$, according to (3). By Lemma 1, $\hat{c}_n(\rho_L) \leq \hat{c}_n(\hat{\rho}_n) = \hat{\sigma}_n^2 / \hat{\rho}_n^{2\nu} \leq \hat{c}_n(\rho_U)$ for all n with probability one. Combining this with Theorem 2 applied to $\hat{c}_n(\rho_L)$ and $\hat{c}_n(\rho_U)$ implies the result. \square

Theorem 3 is useful because it justifies the procedure that is most often adopted in practice, of allowing the range parameter to be estimated from data. Based on arguments following from Zhang (2004), a few authors have advocated for fixing the range parameter in practice. However, as we shall show in Section 3, the estimator $\hat{c}_n(\rho_1)$ can often display sizeable bias, making the approximation in Theorem 2 quite inaccurate. Confidence intervals constructed using Theorem 2 can, due to this bias, have empirical coverage probabilities very near to zero in some cases. In contrast, we will show that confidence intervals for the maximum likelihood estimator, constructed using Theorem 3, have close to nominal coverage even for moderate sample sizes.

2.2. Prediction at New Locations. We now consider the problem of predicting the value of the process at a new location s_0 not in the set of observation locations S_n . Stein (1988, 1990, 1993, 1999) has considered this problem when an incorrect model is used. Predictors under the wrong model can be consistent under relatively weak conditions. Our focus is therefore on two other desirable properties, asymptotic efficiency and asymptotically correct estimation of prediction variance. In a seminal paper, Stein (1988) showed that both of these properties hold when the model used is equivalent to the true measure. In fact, these properties hold uniformly over predictands that are linear functionals of Z or L_2 limits of such functionals, such as integrals of Z (Stein, 1990). In the case of the Matérn covariance, Theorem 1 indicates that this holds for a model with the correct ν and microergodic parameter $\sigma^2/\rho^{2\nu}$. This has led to statements in the literature to the effect that “the parameter $c = \sigma^2/\rho^{2\nu}$ can be consistently estimated, and this is what matters for prediction.” While this statement contains an element of truth, we will argue in this section that it can also be somewhat misleading, both in an asymptotic sense, as well as in guiding choices for applications.

Under the mean zero Gaussian process model with Matérn covariance function and known $\nu > 0$, define

$$(5) \quad \hat{Z}_n(\rho) = \gamma_n(\rho)^T \Gamma_n(\rho)^{-1} Z_n,$$

where $\gamma_n(\rho) = \{K(s_0 - s_i; \rho, \nu)\}_i$ and $\Gamma_n(\rho) = \{K(s_i - s_j; \rho, \nu)\}$ for $i, j = 1, \dots, n$. $\hat{Z}_n(\rho)$ is the best linear unbiased predictor for $Z_0 = Z(s_0)$ under a presumed model $G(0, \sigma^2 K_{\rho, \nu})$ for any value of σ^2 . The predictor itself does not depend on σ^2 , only ρ and ν . Therefore, any intuition that one can fix $\rho = \rho_1$, and that as n increases, c will be better estimated by $\hat{c}_n(\rho_1)$, and for this reason plug-in predictions will improve, is clearly a misunderstanding of asymptotic results. Equivalence, although sufficient for asymptotic

efficiency, is not necessary. The way in which c is relevant for prediction concerns estimates of the mean squared error of the predictor. Under model $G(0, \sigma_0^2 K_{\rho_0, \nu})$, this is

$$(6) \quad \text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho) - Z_0) = \sigma_0^2[1 - 2\gamma_n(\rho)^T \Gamma_n(\rho)^{-1} \gamma_n(\rho_0) + \gamma_n(\rho)^T \Gamma_n(\rho)^{-1} \Gamma_n(\rho_0) \Gamma_n(\rho)^{-1} \gamma_n(\rho)]$$

where $\gamma_n(\rho_0)$ and $\Gamma_n(\rho_0)$ are defined similarly to their counterparts using ρ . In the case that $\rho = \rho_0$, this expression simplifies to

$$(7) \quad \text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_0) - Z_0) = \sigma_0^2[1 - \gamma_n(\rho_0)^T \Gamma_n(\rho_0)^{-1} \gamma_n(\rho_0)].$$

In practice, it is common to estimate the model parameters and then plug them into (5) and (7), treating them as known. The asymptotic properties of this procedure, so-called “plug-in prediction,” are quite difficult to obtain. Instead, most theoretical development has been under a framework in which model parameters are fixed and do not vary with n . We will review these results and indicate how they may be extended to include estimation of the variance parameter σ^2 with a fixed value of ρ , making precise the sense in which the statement regarding c at the beginning of this section should be interpreted. Our simulation results indicate that prediction mean squared error and coverage of prediction intervals can be improved by estimating ρ as well. However, extending the asymptotic results to this case has thus far been intractable. We return to a discussion of this problem at the end of this section.

The following result is an application of Theorems 1 and 2 of [Stein \(1993\)](#).

THEOREM 4. *Suppose $G(0, \sigma_0^2 K_{\rho_0, \nu})$ and $G(0, \sigma_1^2 K_{\rho_1, \nu})$ are two Gaussian process measures on D with the same value of $\nu > 0$.*

1. *As $n \rightarrow \infty$,*

$$\frac{\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_1) - Z_0)}{\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_0) - Z_0)} \rightarrow 1.$$

2. *Furthermore, if $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$, then as $n \rightarrow \infty$,*

$$(8) \quad \frac{\text{Var}_{\sigma_1^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)}{\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_1) - Z_0)} \rightarrow 1.$$

PROOF. Let f_0 be the spectral density corresponding to $\sigma_0^2 K_{\rho_0, \nu}$ and f_1 be the spectral density corresponding to $\sigma_1^2 K_{\rho_1, \nu}$. The result follows from noting

that the function $f_0(\omega)\|\omega\|^{2\nu+d}$ is bounded away from zero and infinity as $\|\omega\| \rightarrow \infty$ and that

$$\lim_{\|\omega\| \rightarrow \infty} \frac{f_1(\omega)}{f_0(\omega)} = \frac{\sigma_1^2/\rho_1^{2\nu}}{\sigma_0^2/\rho_0^{2\nu}}.$$

These two conditions satisfy those needed for Theorems 1 and 2 of [Stein \(1993\)](#). \square

The implication of Theorem 4 is that if the correct value of ν is used, any value of ρ will give asymptotic efficiency. The condition $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$ is not necessary for asymptotic efficiency, but it does provide asymptotically correct estimates of risk. The numerator in (8) is the naive mean squared error for $\hat{Z}_n(\sigma_1^2, \rho_1)$, assuming model $G(0, \sigma_1^2 K_{\rho_1, \nu})$, whereas the denominator is the true mean squared error for $\hat{Z}_n(\sigma_1^2, \rho_1)$, under model $G(0, \sigma_0^2 K_{\rho_0, \nu})$. We now show the same convergence happens if ρ is fixed at ρ_1 but σ^2 is estimated via maximum likelihood. This is an extension of part 2 of Theorem 4. Part 1 needs no extension, since the form of the predictor itself does not depend on σ^2 .

THEOREM 5. *Suppose $G(0, \sigma_0^2 K_{\rho_0, \nu})$ is a Gaussian process measure on D . Fix $\rho_1 > 0$. For a sequence of observations Z_n on an increasing sequence of subsets S_n of D , define $\hat{\sigma}_n^2 = Z_n^T \Gamma_n(\rho_1)^{-1} Z_n/n$. Then as $n \rightarrow \infty$,*

$$(9) \quad \frac{\text{Var}_{\hat{\sigma}_n^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)}{\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_1) - Z_0)} \rightarrow 1$$

almost surely under $G(0, \sigma_0^2 K_{\rho_0, \nu})$.

PROOF. Define $\sigma_1^2 = \sigma_0^2(\rho_1/\rho_0)^{2\nu}$. Then write

$$\frac{\text{Var}_{\hat{\sigma}_n^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)}{\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_1) - Z_0)} = \frac{\text{Var}_{\hat{\sigma}_n^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)}{\text{Var}_{\sigma_1^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)} \frac{\text{Var}_{\sigma_1^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)}{\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_1) - Z_0)}$$

By Theorem 4, $\text{Var}_{\sigma_1^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)/\text{Var}_{\sigma_0^2, \rho_0}(\hat{Z}_n(\rho_1) - Z_0) \rightarrow 1$. So we need only show that $\text{Var}_{\hat{\sigma}_n^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)/\text{Var}_{\sigma_1^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0) \rightarrow 1$ almost surely under $G(0, \sigma_0^2 K_{\rho_0, \nu})$. By (7), $\text{Var}_{\hat{\sigma}_n^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0)/\text{Var}_{\sigma_1^2, \rho_1}(\hat{Z}_n(\rho_1) - Z_0) = \hat{\sigma}_n^2/\sigma_1^2$. Under $G(0, \sigma_1^2 K_{\rho_1, \nu})$, $\hat{\sigma}_n^2$ is equal in distribution to σ_1^2/n times a χ^2 random variable with n degrees of freedom and hence converges almost surely to σ_1^2 as $n \rightarrow \infty$. Because $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$, Theorem 1 gives that $G(0, \sigma_0^2 K_{\rho_0, \nu})$ and $G(0, \sigma_1^2 K_{\rho_1, \nu})$ are equivalent, so that $\hat{\sigma}_n^2 \rightarrow \sigma_1^2$ almost surely under $G(0, \sigma_0^2 K_{\rho_0, \nu})$ as well. \square

We conjecture that the asymptotic behavior in Theorem 4.1 and Theorem 5 still holds if ρ_1 is replaced by $\hat{\rho}_n$, the maximum likelihood estimator, although proving such a result has thus far been intractable. The primary difficulty is that proofs for fixed ρ are able to exploit equivalence, while it is generally not the case that $G(0, \hat{\sigma}_n^2 K_{\hat{\rho}_n, \nu})$ and $G(0, \sigma_0^2 K_{\rho_0, \nu})$ are equivalent for any finite n . Putter and Young (2001) provide conditions under which both asymptotic efficiency and asymptotically correct risk estimation hold for plug-in predictions. They require that the sequence of measures be contiguous. However, they are only able to demonstrate contiguity under extremely limited conditions, including the Matérn covariance with $\nu = 1/2$ when the observations are equally spaced in one dimension. They note that verifying contiguity “in more general (parametric) covariance function models however, such as the Matérn model, as proposed by Stein (1999), may pose formidable problems.”

One approach we have considered is to prove that expression (6) for the actual mean squared error of $\hat{Z}(\rho)$ is quasi-convex in ρ . This would allow a similar method of proof as in Theorem 3. Although we have not been able to find a counter-example to this statement, we have also not been able to prove that it is true. Viewed as a function of the weight vector $\Gamma_n(\rho)^{-1} \gamma_n(\rho)$, (6) is clearly convex. However, the weight vector does not satisfy any simple conditions that could be used to show quasi-convexity when it is viewed as a function of ρ .

Despite these difficulties, we believe it is important to continue studying the full plug-in prediction problem. Simulation results in the next section indicate that a procedure that uses the maximum likelihood estimators can have important advantages over a procedure that fixes the range parameter.

3. Simulation Study. Fixing the range parameter is supported by asymptotic results, and it is computationally efficient in practice, as expensive computations involving the correlation matrix only need to be carried out once. However, it is unclear to what degree asymptotic results are appropriate in guiding our choices for applied problems with finite sample sizes. To systematically explore this issue, we simulate data under a Gaussian process model for a variety of settings chosen to mimic the range of behavior we might observe in practice, and we compare the performance of procedures that either fix or estimate the range parameter. In some cases the performance metrics can be calculated analytically rather than via simulation, as we will indicate.

We simulate data under the mean zero Gaussian process model with Matérn covariance and smoothness parameter $\nu = 0.5$ or 1.5 and marginal

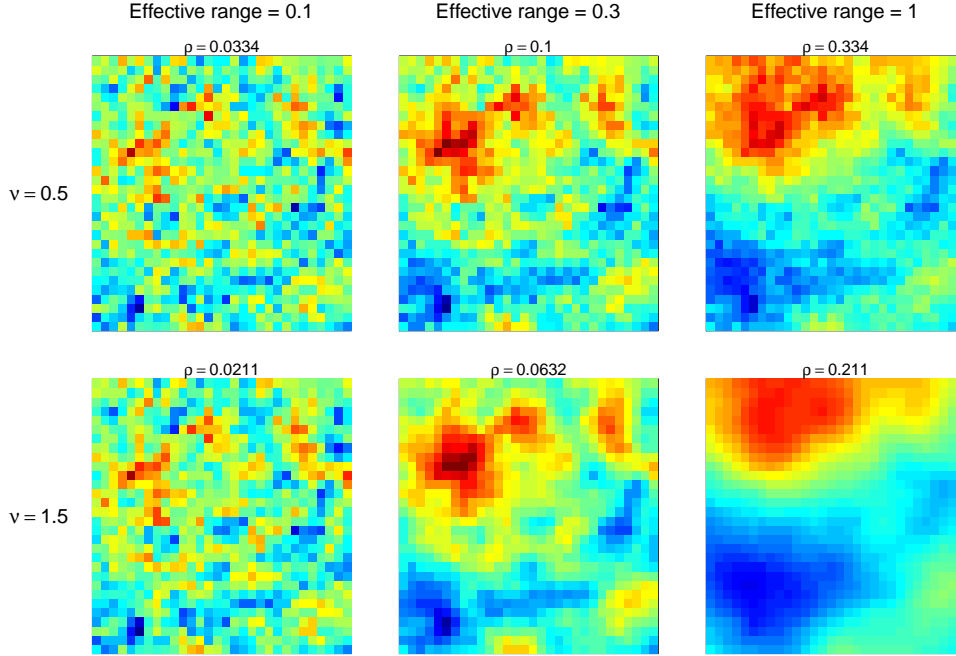


FIG 1. *Simulated random fields on $[0, 1]^2$ under parameter settings used in the simulation study. Fields were simulated using the Cholesky method, using the same set of random deviates for each panel. The value of the range parameter ρ corresponding to each ν and effective range combination is also indicated.*

variance $\sigma^2 = 1$. We also use three effective ranges for the process. That is, we choose values of ρ such that the correlation decays to 0.05 at these prescribed distances. The distances chosen as effective ranges are 0.1, 0.3, and 1. (We also carried out an expanded version of this simulation study incorporating the settings $\nu = 1.5$ and distances 0.05, 0.2, 0.6, and 2. The pattern of results was consistent with what we report here.) Figure 1 illustrates the effect of these parameter settings when simulating from a Gaussian process over the unit square. The datasets in Figure 1 are each generated using the Cholesky decomposition method (Cressie, 1993, Section 3.6.1), transforming the same set of independent standard normal random samples for each parameter setting. As we shall see, whether a particular sample size is “large enough” to be approximated by asymptotic results depends on both the degree of smoothness and the effective range of the process.

In addition to the effective range and smoothness, we also vary the sample size in the simulation. In computing our performance metrics, we use datasets of size $n = 100, 400, 900$, and 1600. To avoid numerical issues that

can arise from sampling locations situated too close to each other, sampling locations are constructed using a perturbed grid. We construct a regular grid with coordinates from 0.005 to 0.995 in increments of 0.015 in each dimension. To each gridpoint, we add a random perturbation according to a uniform distribution over $[-0.005, 0.005]^2$. The resulting set of 4489 locations therefore has the property that all pairs of points have at least 0.005 distance from each other. We then choose random subsets of these locations to be our $n = 100, 400, 900$, or 1600 observation locations, with each sample size containing the points from smaller sample sizes. In evaluating the predictive properties of models fit using a fixed or estimated range parameter, we consider a 50×50 regular grid of locations over $[0, 1]^2$.

For each parameter setting, we simulate 1000 datasets corresponding to realizations of the Gaussian process observed at the union of $n = 1600$ observation and $m = 2500$ prediction locations. For each dataset and sample size, we estimate σ^2 and ρ by numerically maximizing the likelihood. We also calculate $\hat{\sigma}_n^2(\rho_1) = Z_n^T \Gamma_n(\rho_1)^{-1} Z_n / n$ for values of ρ_1 equal to 0.2, 0.5, 1, 2, and 5 times the true value of ρ . Corresponding to each of these parameter estimates, we also construct 95% confidence intervals for $c = \sigma^2 / \rho^{2\nu}$ using the normal approximation provided by Theorem 2 when ρ is fixed and Theorem 3 when ρ is estimated. Finally, we construct kriging predictors and estimated standard errors for each of the $m = 2500$ prediction locations by plugging in parameter estimates into (5) and (7).

The next sections discuss the results for estimation and prediction. Many of the results show a similar pattern, which can be summarized as follows. The performance of the maximum likelihood estimator, maximizing over both σ^2 and ρ , is generally very good, especially by $n = 1600$. Procedures using a fixed ρ are almost always worse, although there are certain settings under which the differences are minimal. These tend to be for $\nu = 1/2$ (corresponding to processes that are not mean-square differentiable) and a large effective range. In these cases, particularly when ρ is fixed at something larger than its true value, the estimators and predictors can still perform well. This agrees with some examples in the literature, for which $\nu = 1/2$ and large effective ranges were used (Zhang and Wang, 2010; Wang and Loh, 2011). When the process is smooth ($\nu = 1.5$) and/or the true range of spatial correlation is small, estimation and prediction is markedly improved by estimating ρ via maximum likelihood.

3.1. Parameter estimation. Given the asymptotic results in Zhang (2004) and Wang and Loh (2011) for $\hat{c}_n(\rho_1)$ for fixed ρ_1 , it is tempting to adopt the intuition that this estimator can “adapt” to incorrectly specified values

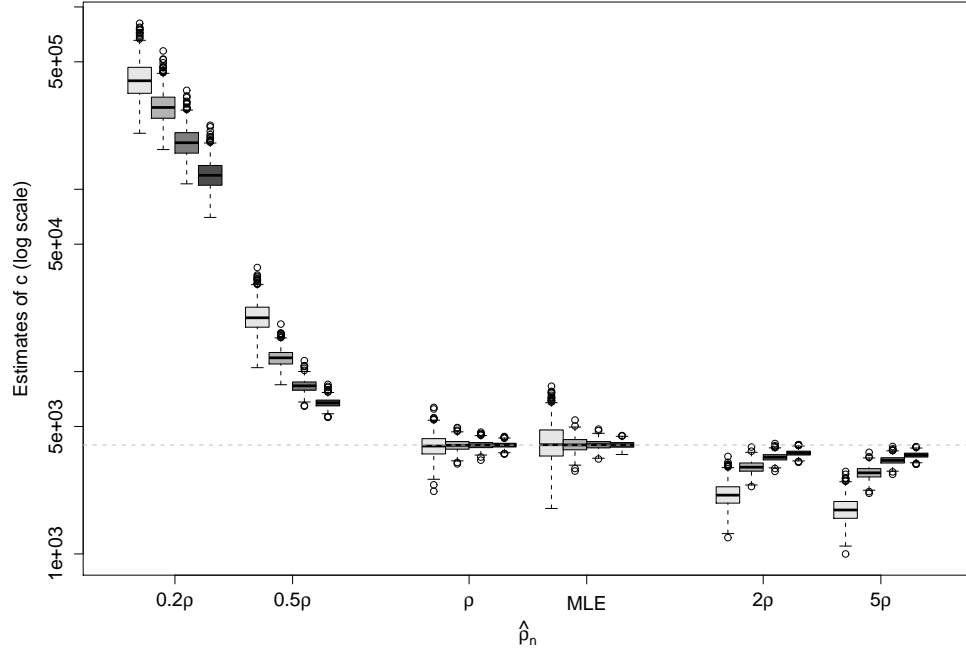


FIG 2. Sampling distributions for \hat{c}_n when $\nu = 1.5$ and the effective range is 0.3. The range parameter is either fixed at the true value (Truth), estimated via maximum likelihood (MLE), or fixed at a multiple of the truth (0.2 Truth, \dots , 5 Truth). The four boxplots in each grouping correspond to sample sizes of $n = 100, 400, 900$, and 1600 , reading from left to right.

of ρ . While this is true asymptotically, we observe in our simulation results that this adaptation in many cases requires a very large value of n ; sampling distributions can be highly biased and can move very slowly toward the truth as n increases. Figure 2 illustrates these effects for a subset of our simulation results, namely when $\nu = 1.5$ and the effective range is 0.3 (This corresponds to the center-bottom panel of Figure 1). Sampling distributions for $\hat{c}_n(\rho_1)$ are noticeably biased. As we expect based on Theorem 2, these biases decrease with n , although even when $n = 1600$ the true value of c lies far in the tail of the sampling distribution. In contrast, the sampling distributions for the maximum likelihood estimator $\hat{c}_n(\hat{\rho}_n)$ have negligible bias. Indeed, they behave very similarly to those for the estimator of c that fixes ρ at the truth.

Similar effects can be seen for other values of ν and effective range. Table 1 shows the relative bias of different estimators of c . Within a given value of ν , the effects are most apparent when the effective range is small. The bias is larger when ν is 1.5 rather than 0.5. Fixing ρ at a value smaller than ρ_0

seems to have a more deleterious effect on bias than fixing ρ larger than ρ_0 . In all cases, including when the sample size n is large, estimating ρ results in notably better performance relative to fixing ρ .

TABLE 1
Bias in $\hat{c}_n(\hat{\rho}_n)$ relative to the true value of c , for the maximum likelihood estimator (MLE) and when $\hat{\rho}_n$ is fixed at a multiple of the true value of ρ . Values labelled as 0.00 should be interpreted to mean less than 0.01.

$\hat{\rho}_n$	n	$\nu = 0.5$			$\nu = 1.5$		
		$er = 0.1$	$er = 0.3$	$er = 1$	$er = 0.1$	$er = 0.3$	$er = 1$
MLE	100	0.60	0.03	0.01	0.18	0.05	0.02
	400	0.01	0.00	0.00	0.02	0.01	0.01
	900	0.00	0.00	0.00	0.01	0.01	0.00
	1600	0.00	0.00	0.00	0.00	0.00	0.00
0.2ρ	100	3.96	3.50	1.61	121.76	101.66	41.49
	400	3.86	2.61	0.64	117.59	71.53	14.17
	900	3.65	1.80	0.33	109.91	45.31	6.73
	1600	3.41	1.27	0.20	101.35	29.70	3.91
0.5ρ	100	0.94	0.63	0.18	6.47	4.11	1.15
	400	0.81	0.33	0.07	5.49	2.02	0.40
	900	0.64	0.19	0.03	4.27	1.12	0.20
	1600	0.51	0.12	0.02	3.29	0.71	0.12
2ρ	100	-0.36	-0.16	-0.05	-0.76	-0.46	-0.16
	400	-0.24	-0.08	-0.02	-0.60	-0.24	-0.06
	900	-0.16	-0.04	-0.01	-0.45	-0.14	-0.03
	1600	-0.12	-0.03	-0.01	-0.35	-0.10	-0.02
5ρ	100	-0.49	-0.21	-0.07	-0.88	-0.56	-0.20
	400	-0.30	-0.10	-0.02	-0.70	-0.30	-0.07
	900	-0.20	-0.05	-0.01	-0.54	-0.18	-0.04
	1600	-0.15	-0.04	-0.01	-0.42	-0.12	-0.02

If Theorem 2 is used to construct confidence intervals and n is in fact not large enough for the normal approximation to be appropriate, the coverage of such intervals can be disastrously low. This is clearly the case for fixed ρ_1 in Figure 2, for which the sampling distributions are extremely biased even when $n = 1600$. Empirical coverage rates for confidence intervals constructed using $\hat{c}_n(\rho_1)$ and $\hat{c}_n(\hat{\rho}_n)$ are shown in Table 2. These were constructed as $\hat{c}_n(\hat{\rho}_n) \pm 1.96 \sqrt{2\hat{c}_n(\hat{\rho}_n)^2/n}$ for $\hat{\rho}_n$ equal to the maximum likelihood estimator or a fixed ρ_1 . Combining parts 1 and 2 of Theorems 2 and 3 implies that these intervals are asymptotically valid 95% confidence intervals for c . Not surprisingly, however, given the large biases observed when ρ is fixed, the differences in the empirical coverage rates between fixed and estimated ρ

are striking, even when n is large. In many cases the coverage for intervals constructed using $\hat{c}_n(\rho_1)$ was 0%, to within Monte Carlo sampling error. The results across different values of ν and effective range mimic the pattern for the bias, with coverage being best when ν is small and effective range is large. For fixed ρ_1 , it also appears better to choose $\rho_1 > \rho_0$ than $\rho_1 < \rho_0$.

Another conclusion we draw from this table is that the sample size needed for the asymptotic approximation in our Theorem 3 to be appropriate is quite a bit smaller than is needed for the equivalent result from Wang and Loh (2011) when ρ is fixed. This does not contradict the simulation results in Wang and Loh (2011). These authors present results indicating that quantiles of the sampling distribution for $\hat{c}_n(\rho_1)$ are well matched by asymptotic approximations, but their data were simulated only for small values of ν (0.25 and 0.5) and large effective ranges (2.8 and 3.7). Intuitively, such a pattern of results is not surprising, because as discussed in Zhang and Zimmerman (2005), larger effective ranges can be thought of as corresponding better to the infill asymptotic framework. However, given that the performance of the maximum likelihood $\hat{c}_n(\hat{\rho}_n)$ is superior across a range of parameter settings, often markedly so, the additional computational expense of maximizing the profile likelihood for ρ seems a worthwhile investment.

3.2. Prediction. The mean squared error of predictor $\hat{Z}_n(\rho_1)$ may be calculated in closed form using (6), rather than relying on simulated data. When the plug-in predictor $\hat{Z}_n(\hat{\rho}_n)$ is used, we need to integrate over the sampling distribution for $\hat{\rho}_n$, which we can do using the simulation results from Section 3.1. For both fixed and estimated ρ , we calculate the average mean squared prediction error, averaging over the $m = 2500$ prediction points in the domain. Because the prediction problem varies in difficulty according to ν , effective range, and sample size n , we report the percent increase in mean squared prediction error relative to the optimal mean squared prediction error using the true value of ρ , which is calculated from (7). These results are shown in Table 3.

Looking at Table 3, it is clear that plug-in prediction using the maximum likelihood estimator $\hat{\rho}_n$ performs quite well relative to predicting with the true value of ρ . For $n = 900$ and 1600, the increase in mean squared error is less than 0.1 percent in all cases. It is also clear that there are cases in which it makes little difference if ρ is fixed at an incorrect value. This is true when the effective range is large ($er = 1$) and ρ_1 is fixed at something larger than the true value. However, there are also cases in which fixing ρ can lead to quite a large loss of efficiency. These effects are magnified when we move from $\nu = 0.5$ to $\nu = 1.5$, suggesting that a misspecified value of

TABLE 2
*Empirical coverage rates of confidence intervals for $c = \sigma^2/\rho^{2\nu}$, expressed as percentages.
 Values labelled as 0 should be interpreted to mean less than 1%.*

$\hat{\rho}_n$	n	$\nu = 0.5$			$\nu = 1.5$		
		$er = 0.1$	$er = 0.3$	$er = 1$	$er = 0.1$	$er = 0.3$	$er = 1$
MLE	100	62	85	93	43	77	91
	400	81	92	94	64	87	94
	900	89	94	94	74	91	94
	1600	90	94	94	81	92	95
0.2ρ	100	0	0	1	0	0	0
	400	0	0	0	0	0	0
	900	0	0	1	0	0	0
	1600	0	0	2	0	0	0
0.5ρ	100	2	16	86	0	0	2
	400	0	4	88	0	0	4
	900	0	7	90	0	0	9
	1600	0	13	92	0	0	18
2ρ	100	8	67	89	0	0	67
	400	3	75	93	0	1	83
	900	3	82	93	0	9	89
	1600	5	84	94	0	17	93
5ρ	100	0	50	86	0	0	55
	400	0	63	92	0	0	77
	900	0	75	93	0	2	86
	1600	0	79	93	0	5	90

ρ is more problematic for smoother processes. This aligns with some earlier cases in the literature in which predictions with a fixed ρ were still quite accurate. For example, [Zhang and Wang \(2010\)](#) examined precipitation data using a predictive process model ([Banerjee et al., 2008](#)) and concluded that a variety of prediction metrics did not change when ρ was fixed at values larger than the maximum likelihood estimator. However, the underlying covariance model for the predictive process was Matérn with $\nu = 1/2$, corresponding to a process that is not mean square differentiable. The model also incorporated a mean term and measurement error, which is beyond the scope of what we consider here.

Table 4 shows the empirical coverage rates of prediction intervals constructed using the naive variance estimator in (7), replacing ρ_0 with either ρ_1 or $\hat{\rho}_n$. In calculating these rates, we average over both simulation replications and over prediction locations. In a similar pattern to what we observe for mean squared error in Table 3, using the maximum likelihood estimator

TABLE 3

Percent increase in mean squared prediction error relative to the optimal mean squared prediction error using the true value of ρ . Values labelled as 0.0 should be interpreted to mean less than 0.1 percent.

$\hat{\rho}_n$	n	$\nu = 0.5$			$\nu = 1.5$		
		$er = 0.1$	$er = 0.3$	$er = 1$	$er = 0.1$	$er = 0.3$	$er = 1$
MLE	100	1.3	0.7	0.5	1.0	0.6	0.8
	400	0.2	0.1	0.0	0.2	0.1	0.1
	900	0.1	0.0	0.0	0.1	0.0	0.0
	1600	0.0	0.0	0.0	0.0	0.0	0.0
0.2ρ	100	11.8	52.3	35.9	24.9	191.3	429.4
	400	36.6	60.4	6.5	103.1	487.0	165.5
	900	56.4	37.5	2.3	218.2	474.1	83.8
	1600	66.2	19.2	0.9	351.4	321.5	41.5
0.5ρ	100	4.3	7.9	1.7	9.8	31.6	20.7
	400	8.7	2.8	0.2	26.9	20.0	2.9
	900	7.9	1.1	0.1	32.9	10.2	1.3
	1600	5.5	0.4	0.0	29.2	4.7	0.7
2ρ	100	5.2	1.8	0.3	15.5	10.2	4.2
	400	2.8	0.3	0.0	12.0	2.1	0.3
	900	1.3	0.1	0.0	6.8	1.0	0.1
	1600	0.6	0.0	0.0	3.4	0.4	0.1
5ρ	100	15.1	4.0	0.8	61.3	29.2	10.3
	400	5.6	0.6	0.1	27.2	4.2	0.6
	900	2.4	0.2	0.0	13.7	1.9	0.2
	1600	1.1	0.1	0.0	6.6	0.9	0.1

produces intervals with the nominal rate in nearly all cases, and the estimators fixing ρ at something larger than the true value achieve this rate for $n = 900$ and 1600 when the effective range is large ($er = 1$). However, the intervals tend to be too conservative when the effective range is large and ρ_1 is too small, and they tend to be not conservative enough when the effective range is small and ρ_1 is too big. Another striking finding is that coverage can sometimes become worse initially as n increases. See for example the entries in Table 4 for which $\nu = 1.5$, $\hat{\rho}_n = 0.2\rho$, and the effective range is 0.3. Although we know from Theorem 4 that this coverage will eventually be close to nominal, the sample size needed for this to occur may be quite large, and coverage can get worse with n before it improves, making it difficult to judge when fixing ρ is “safe” in practice.

TABLE 4
Empirical coverage rates of prediction intervals, expressed as percentages.

$\hat{\rho}_n$	n	$\nu = 0.5$			$\nu = 1.5$		
		$er = 0.1$	$er = 0.3$	$er = 1$	$er = 0.1$	$er = 0.3$	$er = 1$
MLE	100	94	95	95	94	95	95
	400	95	95	95	95	95	95
	900	95	95	95	95	95	95
	1600	95	95	95	95	95	95
0.2ρ	100	95	96	98	95	96	99
	400	96	98	98	96	98	100
	900	96	98	97	97	99	100
	1600	97	99	97	98	100	100
0.5ρ	100	95	96	96	95	97	98
	400	96	97	95	97	99	97
	900	97	96	95	98	99	97
	1600	97	96	95	99	98	96
2ρ	100	93	94	94	90	91	93
	400	93	94	95	90	93	94
	900	94	95	95	91	94	95
	1600	94	95	95	92	94	95
5ρ	100	92	93	94	87	89	93
	400	92	94	95	87	92	94
	900	93	94	95	89	93	95
	1600	94	95	95	90	94	95

4. Covariance tapering. We return now to asymptotic theory in a slightly different but related setting, showing that our approach in Section 2.1 is not limited to the maximum likelihood estimator. The method of covariance tapering has been suggested as a means of coping with computational difficulties posed by large spatial datasets (Furrer, Genton and Nychka, 2006; Kaufman, Schervish and Nychka, 2008). The likelihood function (2) may be replaced by

(10)

$$\mathcal{L}_{n,tap}(\sigma^2, \rho) = (2\pi\sigma^2)^{-n/2} |\Gamma_n(\rho) \circ T_n|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} Z_n^T [\Gamma_n(\rho) \circ T_n]^{-1} Z_n \right\}$$

where T_n is a correlation matrix formed using a correlation or “taper” function that has compact support and $A \circ B$ indicates the direct or Hadamard product between matrices A and B of the same dimension. Because tapering can introduce sparsity, maximizing (10) has computational advantages over maximizing the original likelihood function. For a fixed value of ρ , (10) is maximized by $\hat{\sigma}_{n,tap}^2(\rho) = Z_n^T [\Gamma_n(\rho) \circ T_n]^{-1} Z_n / n$, and the corresponding

estimator of c is

$$(11) \quad \hat{c}_{n,tap}(\rho) = \hat{\sigma}_{n,tap}^2(\rho) / \rho^{2\nu} = Z_n' [\Gamma_n(\rho) \circ T_n]^{-1} Z_n / (n\rho^{2\nu}).$$

Under certain conditions on the taper function, $\hat{c}_{n,tap}(\rho)$ is strongly consistent and has the same asymptotic normal distribution as in Theorem 2 (Kaufman, Schervish and Nychka, 2008; Du, Zhang and Mandrekar, 2009; Wang and Loh, 2011). As was the case for the original likelihood, all of these results considered a fixed value of ρ . However, the same argument used in the proof of Lemma 1 can be used to show that $\hat{c}_{n,tap}(\rho)$ is a monotone function of ρ , allowing us to extend the asymptotic results to the case that (10) is maximized over both σ^2 and ρ .

LEMMA 2. *Let $S_n = \{s_1, s_2, \dots, s_n \in D \subseteq \mathbb{R}^d\}$ denote any set of observation locations in any dimension. Fix $\nu > 0$ and define $\Gamma_n(\rho)$ to be the matrix with entries $K(s_i - s_j; \rho, \nu)$ defined using the Matérn correlation function and T_n to be any positive semi-definite matrix. Then for any $0 < \rho_1 < \rho_2$, $\hat{c}_{n,tap}(\rho_2) \leq \hat{c}_{n,tap}(\rho_1)$ for any vector Z_n .*

PROOF. The proof is similar to that for Lemma 1. Replace the matrix B by $[\rho_2^{2\nu}\Gamma_n(\rho_2) - \rho_1^{2\nu}\Gamma_n(\rho_1)] \circ T_n$ and use the fact that the direct product of two positive semi-definite matrices is again positive semi-definite (Horn and Johnson, 1991, Theorem 5.2.1). \square

This can be used to extend Theorems 1 and 2 of Wang and Loh (2011) to cover the case that ρ is estimated by maximizing the tapered likelihood. (Special cases of Theorems 1 and 2 of Wang and Loh (2011) appear as Theorem 2 of Kaufman, Schervish and Nychka (2008) and Theorem 5.ii of Du, Zhang and Mandrekar (2009).)

THEOREM 6. *Let S_n be an increasing sequence of subsets of D . Suppose $(\sigma_0^2, \rho_0)^T \in (0, \infty) \times [\rho_L, \rho_U]$, for any $0 < \rho_L < \rho_U < \infty$. Let K_{tap} be an isotropic correlation function with support $[-1, 1]^d$ whose spectral density*

$$f_{tap}(w) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega^T x} K_{tap}(x) dx$$

satisfies

$$f_{tap}(w) \leq \frac{M}{(1 + \|\omega\|^2)^{\nu+d/2+\epsilon}}$$

for some constants $\epsilon > \max\{d/4, 1 - \nu\}$ and $M > 0$. Define a sequence of positive taper ranges $\gamma_n = Cn^{-a}$ for $0 < C \leq 1$ and $a \geq 0$. For each n , let T_n be the $n \times n$ matrix with entries $K_{tap}(\|s_i - s_j\|/\gamma_n)$ and let $\hat{\sigma}_{n,tap}^2$ and $\hat{\rho}_{n,tap}$ maximize (10) over $(0, \infty) \times [\rho_L, \rho_U]$.

1. Suppose $0 \leq a < 1/(4\nu + 4\epsilon + 2d)$. Then as $n \rightarrow \infty$,

$$\hat{\sigma}_{n,tap}^2 / \hat{\rho}_{n,tap}^{2\nu} \rightarrow \sigma_0^2 / \rho_0^{2\nu}$$

almost surely under $G(0, \sigma_0^2 K_{\rho_0, \nu})$.

2. Suppose $0 \leq a < 1/(8\nu + 8\epsilon + 2)$ and $2a(2\nu + 2\epsilon + d) / \min\{2, 4 - d, 4\epsilon - d, 4\nu + d\} < (1 - 2ad)/(2d)$. Then as $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\sigma}_{n,tap}^2 / \hat{\rho}_{n,tap}^{2\nu} - \sigma_0^2 / \rho_0^{2\nu}) \rightarrow N(0, 2(\sigma_0^2 / \rho_0^{2\nu})^2)$$

in distribution under $G(0, \sigma_0^2 K_{\rho_0, \nu})$.

PROOF. By assumption, $\rho_L < \hat{\rho}_{n,tap} < \rho_U$ for every n . Define two sequences, $\hat{c}_{n,tap}(\rho_L)$ and $\hat{c}_{n,tap}(\rho_U)$, according to (11). By Lemma 2, $\hat{c}_{n,tap}(\rho_L) \leq \hat{c}_{n,tap}(\hat{\rho}_{n,tap}) = \hat{\sigma}_n^2 / \hat{\rho}_{n,tap}^{2\nu} \leq \hat{c}_{n,tap}(\rho_U)$ for all n with probability one. Combining this with Theorems 1 and 2 of Wang and Loh (2011) applied to $\hat{c}_{n,tap}(\rho_L)$ and $\hat{c}_{n,tap}(\rho_U)$ implies the result. \square

REMARK 1. The case $a = 0$ corresponds to a constant taper range C , for which the conditions on a are trivially satisfied.

Simulation results in Kaufman (2006) indicate that the estimator $\hat{\sigma}_{n,tap}^2 / \hat{\rho}_{n,tap}^{2\nu}$ can be noticeably biased in finite samples. This is in contrast to the results for the maximum likelihood estimator shown in Section 3. Our understanding of this result is that (10) essentially imposes a shorter effective range through the taper function. Therefore, the same caution we have urged in interpreting results that fix the range parameter should be applied to Theorem 6 as well. Also, the simulation results in Kaufman (2006) were for a fixed taper range ($a = 0$). The bias is likely to decrease even more slowly with n if the taper range is allowed to decrease as n increases. Kaufman, Schervish and Nychka (2008) introduced another approximation using tapering that does not suffer from this bias, although existing asymptotic results for it do not cover the infill case (Shaby and Ruppert, 2011).

5. Discussion. We have argued for the importance of estimating the range parameter in geostatistical models, even when asymptotic results can be obtained while holding it fixed. We have shown for the widely-used Matérn model that the same asymptotic behavior can be obtained for an estimator of the consistently estimable parameter $c = \sigma^2 / \rho^{2\nu}$ when the variance parameter σ^2 and range parameter ρ are jointly estimated via maximum likelihood, rather than using a fixed value of ρ . As demonstrated in our

simulation study, this can lead to dramatic reductions in bias and improvements in empirical coverage rates of confidence intervals for this parameter. We also clarified the conditions for asymptotic efficiency of predictions under this model, highlighting a common misunderstanding of existing results that fix the range parameter. We extended asymptotic results for prediction to allow the variance parameter to be estimated, although results for plug-in prediction using the maximum likelihood estimator for both σ^2 and ρ are still elusive. However, simulation results support the importance of estimating the range parameter, both in estimating c and in carrying out plug-in prediction. This is particularly evident when the process is smooth (larger ν) or the effective range of the process is small.

We have made a number of simplifying assumptions. Considering the ways in which these assumptions may be relaxed provides a rich set of questions for future research. For example, our results concern only mean zero Gaussian processes, which is equivalent to assuming that the mean of the process is known. Results on equivalence of mean zero Gaussian measures such as Theorem 1 (Zhang, 2004) can be used in proving equivalence of Gaussian process measures with different means (Stein, 1999, Chapter 4, Corollary 5). Demonstrating this basically requires one to show that the difference between mean terms be sufficiently smooth (Yadrenko, 1983, page 138). However, the primary difficulty is in extending estimation results. Zhang (2004) indicates that his method of proof is not easily extended to the case of an unknown mean term. Asymptotic results for the case $\nu = 1/2$ and $d = 1$ are given in Theorem 3 of Ying (1991), and it seems plausible that similar results might hold for $d = 2$ and 3. A more direct route than the method of proof used in (Zhang, 2004) might be to characterize the distribution of the quadratic form in the expression for the maximum likelihood estimator $\hat{\sigma}_n^2$.

We have also not considered what happens when the observations are not of the process Z itself, but of Z observed with measurement error. Again, results for equivalence and prediction can be extended in a relatively straightforward way. We expect something like Theorem 3 should hold for the case that Z is observed with measurement error. However, in a restricted version of this problem, the rate of convergence is known to change when measurement error is also included. Specifically, Chen, Simpson and Ying (2000) extended the results of Ying (1991) for the case $\nu = 1/2$ and $d = 1$ (the Ornstein-Uhlenbeck) process to include the presence of a measurement error term. The introduction of the error term reduces the rate of convergence of the maximum likelihood estimator for c from the usual order $n^{-1/2}$ to order $n^{-1/4}$.

Perhaps the most important restriction, both here and in previous work,

is that the smoothness parameter ν is assumed to be known. Estimating ν provides desirable flexibility, as this parameter controls the mean square differentiability of the process. However, we know of no results concerning the maximum likelihood estimator in this case. Stein (1999, Section 6.7) examines a periodic version of the Matérn model and argues that $\hat{\sigma}_n^2$ and $\hat{\nu}_n$ should have a joint asymptotic normal distribution, but it is an open question whether a similar result holds for non-periodic fields. Proving such a result will likely require a new approach, as the method of proof for showing equivalence used by Zhang (2004) cannot be extended to the case $\sigma_0^2/\rho_0^{2\nu_0} = \sigma_1^2/\rho_1^{2\nu_1}$.

References.

- ABROMOWITZ, M. and STEGUN, I., eds. (1967). *Handbook of Mathematical Functions*. U.S. Government Printing Office.
- ANDERES, E. (2010). On the consistent separation of scale and variance for Gaussian random fields. *The Annals of Statistics* **38** 870–893.
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC.
- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 825–848.
- CHEN, H., SIMPSON, D. and YING, Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica* **10** 141–156.
- CRESSIE, N. A. C. (1993). *Statistics for spatial data*. John Wiley & Sons.
- DU, J., ZHANG, H. and MANDREKAR, V. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *The Annals of Statistics* **37** 3330–3361.
- FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15** 502–523.
- GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association* **105** 1167–1177.
- HORN, R. A. and JOHNSON, C. R. (1985). *Matrix analysis*. Cambridge University Press, Cambridge.
- HORN, R. A. and JOHNSON, C. R. (1991). *Topics in matrix analysis*. Cambridge University Press, Cambridge.
- KAUFMAN, C. G. (2006). Covariance tapering for likelihood-based estimation in large spatial data sets PhD thesis, Carnegie Mellon University.
- KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103** 1545–1555.
- MARDIA, K. and MARSHALL, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146.
- PUTTER, H. and YOUNG, G. (2001). On the effect of covariance function estimation on the accuracy of kriging predictors. *Bernoulli* **7** 421–438.
- SHABY, B. and RUPPERT, D. (2011). Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics*. To appear.

- STEIN, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics* **16** 53–63.
- STEIN, M. L. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *The Annals of Statistics* **18** 850–872.
- STEIN, M. L. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & probability letters* **17** 399–404.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some theory for kriging*. Springer Series in Statistics. Springer-Verlag, New York.
- STEIN, M. L. (2010). Asymptotics for Spatial Processes. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp, eds.) 79–88. CRC.
- WANG, D. and LOH, W. L. (2011). On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electronic Journal of Statistics* **5** 238–269.
- YADRENKO, M. I. (1983). *Spectral theory of random fields*. Optimization Software, Inc.
- YING, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *J. Multivariate Anal.* **36** 280–296.
- ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99** 250–261.
- ZHANG, H. and WANG, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics* **21** 290–304.
- ZHANG, H. and ZIMMERMAN, D. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92** 921–936.

C.G. KAUFMAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: cgk@stat.berkeley.edu

B.A. SHABY
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: bs128@stat.duke.edu